Enhanced Web Mining Technique To Clean Web Log File

Rachit Goel
Department of Computer Science and Engineering
M.tech Scholar, Doon Valley, Karnal

ABSTRACT
The arrival of the computer technology has contributed the ability to produce and store the massive amounts of data. Now the world is not confined only to manually generated files or reports, but has become a giant store where vast amounts of data are collected and exchanged daily.
Web pages typically contain a large amount of information that is not part of the main content of the pages, e.g. banner ads, navigation bars, copyright notices, etc. Such noise on web pages usually leads to poor results in Web Mining which mainly depends upon the web page content. Therefore, it becomes very essential to extract information from the bulks of data and structure them into useful knowledge that will be helpful for some type of understanding. This leads to the birth of data mining. Web usage mining is the subject field of Data Mining which deals with the discovery and analysis of usage patterns from web data specifically web logs in order to improve the web based applications. The motive of mining is to find users' access models automatically and quickly from the vast Web log data, such as frequent access paths, frequent access page groups and user clustering. Through web usage mining, the server log, registration information and other relative information left by user provide foundation for decision making of organizations.

Keywords
Data Usage Mining, Data Preprocessing, Pattern Discovery, Pattern Analysis.

1. INTRODUCTION
The rapid advancement in networking has enabled global telecommunication networks to carry tons of data traffic on a daily basis. With the advent of Internet the World Wide Web has became an important place for information dissemination. There are billions of web search engines (such as Google, Yahoo, Bing, etc.), that processes large amount of data. Today the social media (e.g. Facebook, Orkut, Twitter, etc.) has become an essential element in people's life and hence an important source of data of producing numerous pictures, videos, blogs, etc. The list of sources that generate huge amounts of data is endless. However, the abundant information on the web is not stored in any systematically structured way, which poses great challenges to those looking for high quality information underlying in web pages. The growth in the mass of data present on web has engrossed the attention of the scholars and researchers towards the application of data mining techniques on the data available on the web in order to extract useful information. Web mining has therefore become an important subject matter in data mining. Therefore, it becomes very essential to extract useful information from the bulks of data. According to Frawley, data mining is defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [1]. Data mining can also be defined as the process of discovering meaningful new correlation, patterns and trends by analyzing the large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques. Data mining is a step that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data [2]. Web usage mining is process of discovering usage patterns from Web data, in order to better understand the needs of web based applications. Web usage mining deals with the extraction of knowledge from web server log files. The main source of data for Web usage mining mainly consists of the (textual) logs, which are collected when users access web servers. A high level web usage mining process is shown in figure 1. The Web usage mining is parsed into three distinctive phases [3]:

a) **Data Preprocessing-** It performs a series of steps covering data cleaning, user identification, session identification, path completion and transaction identification.

b) **Pattern Discovery-** It involves application of various data mining techniques to processed data like statistical analysis, association, clustering and pattern matching.

c) **Pattern Analysis-** Filters out irrelevant patterns from the identified patterns generated in pattern discovery phase.

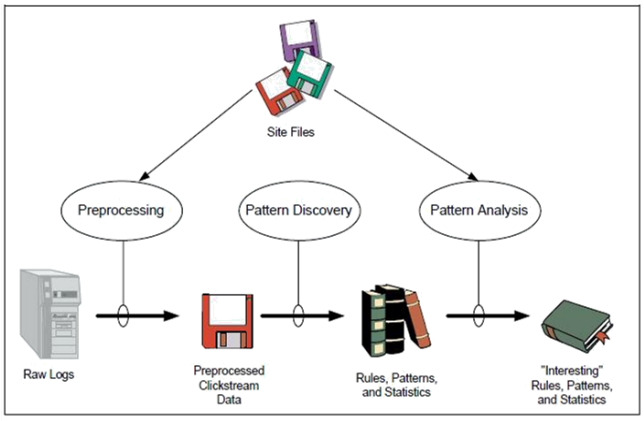

Fig. 1: Overview of Web Usage Mining

2. CHALLENGES ASSOCIATED WITH THE WORLD WIDE WEB
The advent of the World Wide Web has caused a dramatic increase in the usage of the Internet. The World Wide Web is a broadcast medium where a wide range of information can be obtained at a low cost. It has profoundly influenced many aspects of human lives and changed the way of communication, conducting businesses, etc. Additionally the web data is also gigantic, diverse and dynamic in nature due to which users could encounter problems while interacting with the web. The resulting growth in on-line

information combined with the almost unstructured web data necessitates the development of powerful yet computationally efficient web data mining tools. Following are the challenges associated with the WWW:

- **Finding Relevant Information:** The amount of information available on the web is vast and still emergent. Moreover, coverage of the information is very broad and diverse.
- **Heterogeneity:** Information available on Web is heterogeneous in nature. This makes the integration of information from different pages a problem.
- **Noisy:** Information on web is noisy as there is no control over the quality of the information available on the web because there is no restriction on what one writes.
- **Dynamic:** Web is very dynamic in nature since the information available on it changes frequently. Therefore, it is needed to maintain these changes.
- **Redundancy:** Data available on web may be redundant since same segment of information or its variants may appear in many pages.
- **Personalization of Information:** Users of internet differs in their experience intended for the contents they search for and the presentation of their search result.

3. WEB MINING

World Wide Web is a monolithic repository of web pages that provides the Internet users with heaps of information. With the brisk growth of the World Wide Web, the web has become an imperative medium of information dissemination. Therefore, the information available on the Web is a vital source of information for the users of the internet. Due to these reasons there is an increase in the number and size of websites available on internet which makes the World Wide Web remarkably gigantic. However, the abundant information on the web is not stored in any systematically structured way, which poses great challenges to those looking for high quality information underlying in web pages. The resulting growth in on-line information combined with the almost unstructured web data necessitates the development of powerful yet computationally efficient web data mining tools. Oren Etzioni first proposed the term of Web mining in 1996. He claimed the Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services. There were two different approaches proposed for defining Web mining. First approach is a process centric view and second approach is a data centric view. The data centric definition has become more acceptable [10] [11].

i. From the process centric view, web mining is defined as a sequence of ordered tasks [12].

ii. From the data-centric view, web mining is defined with respect to the types of web data that was used in the mining process [7].

Though web mining takes its foundations from the data mining techniques but it does not solely depends on data mining. The reason is that data mining is applied on structured data while web mining is applied on web data which is heterogeneous and unstructured or semi structured in nature. Data mining is work upon offline whereas web mining is work upon online.

4. WEB MINING TASKS

Web mining can be decomposed into the following subtasks [13] [14]:

a) Information Retrieval (or Resource Discovery): Search is probably the one of the prime application of the Web having its roots in information retrieval. Information retrieval (IR) helps the users to find the required information available from a large collection of text documents.

b) Information Extraction (Selection and Preprocessing): This task deals with the transformation of the data retrieved during information retrieval process into a form that can be easily analyzed [17]. Information extraction aims to select relevant facts from the documents while information retrieval aims to select relevant documents.

c) **Generalization (Pattern Recognition and Machine Learning):** It automatically generates general patterns from both the individual web sites as well as across multiple sites. Machine learning methods or data mining techniques are generally used for the generalization purpose.

d) Analysis (Validation and Interpretation): Once the patterns have been identified it is necessary to explore and confirm those mined pattern. The aim of this task is to validate the mined patterns.

Based on the above mentioned subtasks, web mining can be viewed as the use of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web documents and services.

5. WEB MINING TYPES

The advent of the World Wide Web has made the data present on the web as a gigantic source of information. The World Wide Web, having over 350 million pages, continues to grow rapidly at a million pages per day [4]. This increase in the mass of data has turned researcher's attention towards the use of data mining techniques to extract useful information from the web data.

Web mining can be classified into three types [5] as shown in figure 2:

a) Web Content Mining

b) Web Structure Mining

c) Web Usage Mining.

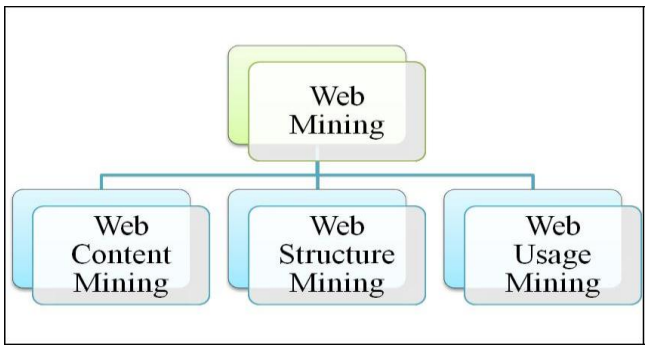

Fig. 2: Web Mining Types

a) **Web Content Mining:** Web content mining describes the automatic search of information resource available online and involves mining web data contents. In the Web mining domain, Web content mining essentially is an analog of data mining techniques for relational databases, since it is possible to find similar types of knowledge from the unstructured data residing in Web documents. Some of them are semi-structured such as HTML documents or a more structured data like the data in the tables or database generated HTML pages, but most of the data is unstructured text data.

b) **Web Structure Mining:** Web structure mining is the process of using graph and network mining theories to comprehend the nodes and hyperlink structures on the Web The goal of Web structure mining is to generate structural summary about the Web site and Web page. Technically, Web content mining mainly focuses on the structure of inner-document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level.

Web Usage Mining: Web usage mining focuses on the extraction of useful information from server logs. Web usage mining tries to discover the useful information from the secondary data derived from the interactions of the users while surfing on the Web. It focuses on the techniques that could predict user behavior while the user interacts with Web.

6. TYPES OF DATA USED FOR MINING

Today there is a myriad collection of data ranging from simple mathematical measurements and text documents to more complex information such as spatial data, multimedia channels, and web documents. Data mining can be applied to any kind of data repository from flat files to databases as long as data is meaningful and useful for the application. However, the approach may differ from data to data. The type of data on which data mining can be applied is the following subsection.

6.1 Relational Databases

A relational database is a collection of data items structured as a set of tables from which data can be accessed. Tables have columns and rows where columns represent attributes and rows represent tuples. Each column contains one or more data categories while each row contains a unique instance of data for the categories defined by the columns. Relational data can be accessed by database query language such as SQL query. Data mining takes advantages from SQL for selection, transformation and consolidation.

6.2 Data warehouse

A data warehouse is a heterogeneous repository of data collected from multiple data sources into a single accessible format. It usually contains historical data derived from transaction data. Data warehouses are constructed through a series of steps involving data cleaning, data integration, data transformation, data loading and data refreshing.

6.3 Spatial Databases

Spatial data provides information about the physical location and shape of geometric objects. Spatial data are in the form of graphic primitives that are in form of points, lines, polygons or pixels. A spatial database is a database optimized to store and query data.

6.4 Multimedia Databases

A multimedia is a combination of different media (i.e., text, pictures, sounds, video, animations, etc.) used to present multimodal information in conjunction with computer technology. Therefore, a multimedia database consists of data types such text, graphics, images, animations, videos, audios, etc.

6.5 Temporal Databases

The temporal data changes over time and it stores history of how data changed over time. A temporal database has associated with it the built-in time aspects.

6.6 Web data

World Wide Web is one of the most interactive and popular medium for dissemination of the information today. World Wide Web organizes data in the form interrelated documents, which are also known as web pages in web terminology.

7. WEB USAGE MINING PROCESS

A detailed web usage mining process with its sub phases is given in figure 3. The three steps involved in Web usage mining process are as follows:

a) **Data Preprocessing**- It performs a series of steps covering:

- Data Cleaning
- Session Identification

b) **Pattern Discovery**- It involves application of various data mining techniques to processed data like

- Statistical Analysis
- Pattern Matching

c) **Pattern Analysis**- It filters out the irrelevant patterns from the identified patterns generated in pattern discovery phase.

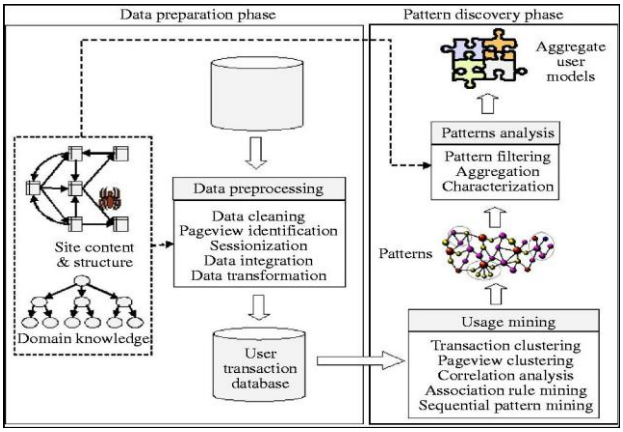

Fig. 3: Detailed Web Usage Mining Process

7.1 Data Preprocessing

The information that can be accessed through web is heterogeneous and semi structured or unstructured in nature. Due to this heterogeneity a web log file may consists of some undesirable log entries whose presence does not matters from the web usage mining point of view. This makes the preprocessing of log file an important precondition for discovering the knowledgeable patterns. The purpose of performing preprocessing is to transform the raw click stream data into a set of user profiles.

Preprocessing enables to translate the unprocessed data which is composed from server log files into constructive data abstraction. Data preprocessing consists of four sub-phases: data cleaning, user identification, session identification and path completion. The usage preprocessing architecture is shown in figure 4.

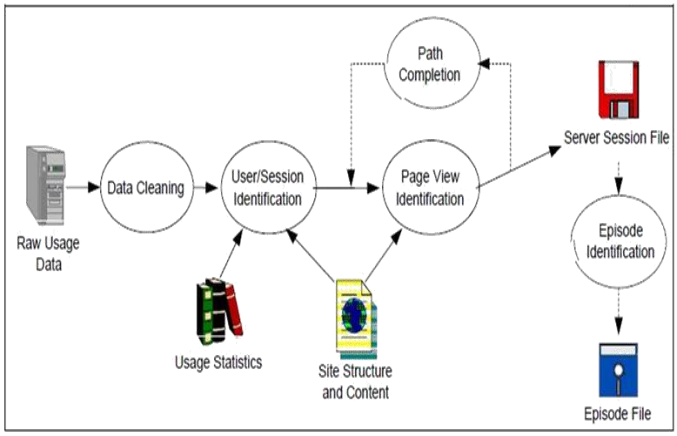

Fig. 4: Usage Preprocessing Architecture

7.2 Pattern Discovery

The second phase of web usage mining is pattern discovery which is the key component of the web mining. Pattern discovery utilizes the algorithms and techniques from several research areas such as data mining, machine learning, statistics and pattern recognition [8].

The methods of pattern discovery are as follows:

a) Statistical Analysis: Statistical techniques are the most common method to extract knowledge about visitors of a web site. By analyzing the session file, one can perform different kinds of descriptive statistical analyses (frequency, mean, median, etc.) on variables such as page views, viewing time and length of a navigational path [9].

b) Association Rules: In the context of Web Usage Mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. These pages may not be directly connected to one another via hyperlinks.

c) Classification: Classification maps a data item into one of several predefined classes. In the Web domain, this technique is used to establish a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class.

7.3 Pattern Analysis

Pattern Analysis is the final stage of the Web usage mining. The goal of pattern analysis is to eliminate the irrelevant rules or patterns from the output of the pattern discovery process. There are two most common approaches for the pattern analysis [9]:

a) SQL query mechanism

b) Construction of multidimensional data cube to perform OLAP operations

8. CONCLUSION

1. The web cleaning step of data preprocessing is crucial as the result of this step have an impact on the accuracy of results of the later phases.

2. An improvised technique for performing the data cleaning technique on server log was proposed.

3. The proposed approach showed a quite salient reduction in the number of records and in the log files size and hence increases the quality of the available data.

9. FUTURE SCOPE

1. The research presented in this paper is in an emerging stage. However, the subjective interpretation of the technique is very ingenious and can propose a lot of scope to be extended on to other problem domains.

2. The research can be extended to the log file of other formats such as extended common log format which consists of more fields than a common log format.

3. Many problems such as applications of user identification, session identification, and path completion are not discussed.

10. REFERENCES

[1] Frawley W.J., Piatetsky-Shapiro G. and Matheus C.J., "Knowledge Discovery in Databases: An Overview", *AI Magazine*, vol. 13, no. 3, pp. 57-70, 1992. [2] Kloesgen, W. 1996. A Multipattern and Multistrategy

Discovery Assistant. In Advances in Knowledge

Discovery and Data Mining.

[3] Srivastava J., and Cooley R., "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM *SIGKDD Explorations*, vol. 1, no. 2, pp. 12-23, January 2000.

[4] Bharat K. and Broder A., "A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines", *in Proceedings of the 7th World-Wide Web Conference*, pp. 379-388, 1998.

[5] Singh B., Singh H.K., "Web Data Mining Research", in *Proceedings of 2010 IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1-10, December 2010.

[6] Bayir M.A., Toroslu I.H., Cosar A. and Fidan G.,"Smart Miner: A New Framework for Mining Large Scale Web Usage Data", in *Proceedings of the 18th International Conference on World Wide Web*, pp. 161-170, 2009.

[7] Cooley R., "Web Usage Mining: Discovery and Application of Interesting Patterns from Web data", PhD thesis, University of Minnesota, Dept. of Computer Science, May 2000.

[8] Singh B., Singh H.K., "Web Data Mining Research", in *Proceedings of 2010 IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1-10, December 2010.

[9] Zhang Q., and Segall R. S., "Web Mining: A Survey of Current Research, Techniques, and Software", *International Journal of Information Technology & Decision Making*, vol. 7, no. 4, pp. 683-720,2008.

[10] Borges J. and Levene M., "Data Mining of User Navigation Patterns", in *Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profiling*, pp. 31-39, August 1999.

[11] Madria S.K., Bhowmick S.S., Ng W.K., and Lim E.P., "Research Issues in Web data Mining", in *Proceedings of First International Conference Data Warehousing and Knowledge Discovery*, pp. 303-312, 1999.

[12] Etzioni O., "The World Wide Web: Quagmire or Gold Mining?", *Communications of the ACM*, vol. 39, no. 11, pp. 65-68, November 1996.

[13] Blockeel H. and Kosala R., "Web Mining Research: A Survey", *ACM SIGKDD Explorations*, vol. 2, no. 1, pp. 1-15, June 2000.

[14] Codd E.F., "A Relational Model of Data for Large Shared Data Banks", *Communications of the ACM*, vol. 13, no. 6, pp. 377–387, June 19.